# ACE 564
# Spring 2006

## *Lecture 11*

## *Violations of Basic Assumptions IV: Specification Errors*

**by**
**Professor Scott H. Irwin**

## Readings:

**Griffiths, Hill and Judge. "Statistical Implications of Misspecifying the Set of Regressors," Section 9.5; "Selecting Regressor Variables and Model Misspecification," Section 10.9 and "Choice of Functional Form," Section 10.10 in *Learning and Practicing Econometrics***

**Kennedy. "The Classical Regression Model," Chapter 3; "Specification," Chapter 5; "Violating Assumption One: Wrong Regressors, Nonlinearities and Parameter Inconstancy," Chapter 6 in *A Guide to Econometrics***

## Specification Errors

Up to this point, in both ACE 562 and 564, we have assumed that the specification of the statistical model is given

In other words, a researcher "knows" the

- Correct choice of independent variables to be included in the model
- Correct choice of functional form
- Correct specification of the error term and its properties

Given specification of the "correct" model, we investigated

- The best way to estimate the model parameters
- How to construct interval estimates of model parameters
- How to test hypotheses about model parameters
- How to forecast based on estimated parameters

Given that all of the major issues we have investigated require knowledge of the correct model, it is natural to ask what happens if we are uncertain about the correct model

- Economic theory often provides only a general guideline regarding the variables to include or exclude in a statistical model
- It may be difficult to find data on some theoretically relevant variables
- Competing economic theories may yield conflicting model specifications
- Uncertainty about the number and form of variables that should appear in a model is one of the most difficult problems facing an applied researcher
- Specification issues have dominated much of the econometrics literature over the last 20 years

Example: explaining product sales by a supermarket

- Almost certain that the price of the product should be included
- Which substitute or complementary product prices should be included?

Example: explaining supply response of hog producers

- Again, almost certain that the price of hogs should be included
- How many lags of hog prices should be included to account for the time between production decisions and the market availability of hog supplies?

In this lecture we are going to assess two important questions regarding model specification

- What are the consequences of choosing the wrong model?
- Are there ways of assessing whether a model is adequate?

## Omitted and Irrelevant Variables

Even with sound economic principles and logic, it is possible that a model specified by a researcher may omit relevant variables or include irrelevant variables

Let's return to the Bay Area Rapid Food model to explore the consequences of an omitted variable

As before, assume the true model is

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$$

where $y_t$ is total chain revenue for week $t$, $x_{2,t}$ is average price of chain products in week $t$, and $x_{3,t}$ is advertising expenditures for week $t$

Suppose that data on advertising expenditures are not available, and as a consequence, we estimate the model using price as the only explanatory variable

$$y_t = \beta_1 + \beta_2 x_{2,t} + e_t$$

By estimating this model, we are imposing the restriction $\beta_3 = 0$ when it is not true

The least squares estimators for $\beta_1$ and $\beta_2$ will generally be biased, although they will have lower sampling variances

The nature of the bias in the LS estimator of $\beta_2$ is especially relevant, so we will derive it rigorously

To begin, remember that the true model is $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$ but we estimate $y_t = \beta_1 + \beta_2 x_{2,t} + e_t$, omitting $x_{3,t}$ from the model

Consequently, we use the "incorrect" simple regression LS estimator

$$b_2^* = \frac{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)(y_t - \bar{y})}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2} = \frac{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)y_t - \sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)\bar{y}}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2}$$

$$= \frac{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)y_t - \bar{y}\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2} = \frac{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)y_t}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2}$$

Now substitute the correct specification of the model into the "incorrect" estimator

$$b_2^* = \frac{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)(\beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t)}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2}$$

which can be re-written as

$$b_2^* = \beta_1 \sum_{t=1}^{T} w_t + \beta_2 \sum_{t=1}^{T} w_t x_{2,t} + \beta_3 \sum_{t=1}^{T} w_t x_{3,t} + \sum_{t=1}^{T} w_t e_t$$

where

$$w_t = \frac{x_{2,t} - \bar{x}_2}{\sum_{t=1}^{T}(x_{2,t} - \bar{x}_2)^2}$$

We can further simplify the estimator to

$$b_2^* = \beta_2 + \beta_3 \sum_{t=1}^{T} w_t x_{3,t} + \sum_{t=1}^{T} w_t e_t$$

because

$$\sum_{t=1}^{T} w_t = 0 \qquad \sum_{t=1}^{T} w_t x_{2,t} = \frac{\sum_{t=1}^{T} (x_{2,t} - \bar{x}_2) x_{2,t}}{\sum_{t=1}^{T} (x_{2,t} - \bar{x}_2)^2} = 1$$

Now take the expected value of the estimator

$$E(b_2^*) = \beta_2 + \beta_3 \sum_{t=1}^{T} w_t x_{3,t} + \sum_{t=1}^{T} w_t E(e_t)$$

Since the expected value of the error term is zero

$$E(b_2^*) = \beta_2 + \beta_3 \sum_{t=1}^{T} w_t x_{3,t} \neq \beta_2$$

We can derive a more meaningful expression for the estimation bias by noting that

$$\sum_{t=1}^{T} w_t x_{3,t} = \frac{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)x_{3,t}}{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)^2} = \frac{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)x_{3,t} - \bar{x}_3\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)}{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)^2}$$

or,

$$\sum_{t=1}^{T} w_t x_{3,t} = \frac{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)(x_{3,t}-\bar{x}_3)}{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)^2}$$

$$\sum_{t=1}^{T} w_t x_{3,t} = \frac{\frac{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)(x_{3,t}-\bar{x}_3)}{T-1}}{\frac{\sum_{t=1}^{T}(x_{2,t}-\bar{x}_2)^2}{T-1}} = \frac{\hat{\text{cov}}(x_{2,t}, x_{3,t})}{\hat{\text{var}}(x_{2,t})}$$

Consequently, we can write the expected value of the omitted variable slope estimator as

$$E(b_2^*) = \beta_2 + \beta_3 \frac{\hat{\text{cov}}(x_{2,t}, x_{3,t})}{\hat{\text{var}}(x_{2,t})} \neq \beta_2$$

Now, note that

$$b_{23} = \frac{\hat{\text{cov}}(x_{2,t}, x_{3,t})}{\hat{\text{var}}(x_{2,t})}$$

is the slope estimator from a regression of $x_{2,t}$ on $x_{3,t}$

As we learned in Lecture 8, such regressions are called "auxiliary regressions" since they have the form of a regression model but are only descriptive devices

The final form of the expected value expression is then

$$E(b_2^*) = \beta_2 + \beta_3 b_{23} \neq \beta_2$$

We can now define the bias in the omitted variable slope estimator as

$$\text{Bias} = E(b_2^*) - \beta_2 = \beta_3 b_{23}$$

Knowing the sign of $\beta_3$, the slope parameter for the omitted variable, and the sign of the correlation between $x_{2,t}$ and $x_{3,t}$ tells us the direction of the bias:

*Tendency to over-estimate* $\beta_2$

$[\beta_3 > 0 \text{ and } b_{23} > 0]$ or $[\beta_3 < 0 \text{ and } b_{23} < 0]$

*Tendency to under-estimate* $\beta_2$

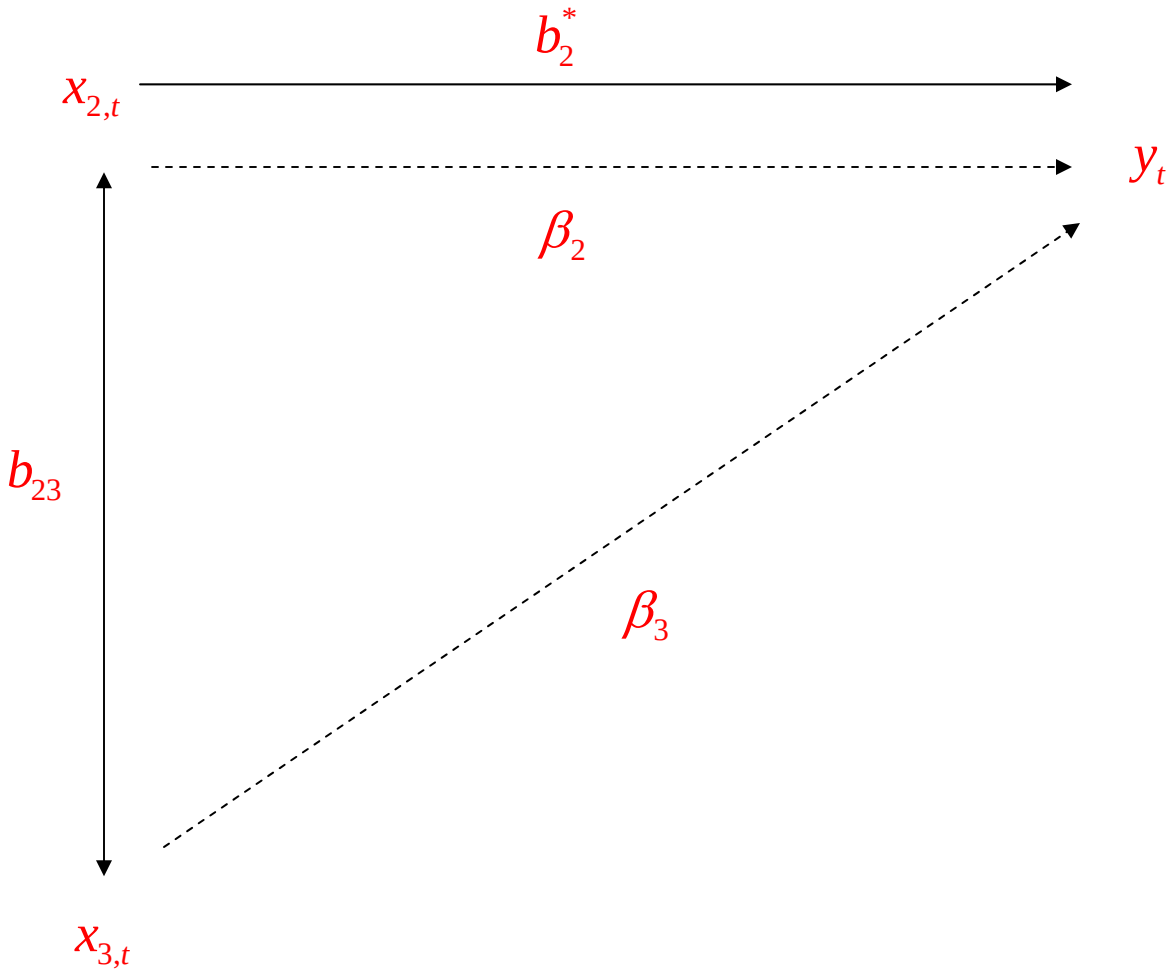
$[\beta_3 > 0 \text{ and } b_{23} < 0]$ or $[\beta_3 < 0 \text{ and } b_{23} > 0]$

One way to think of the bias is that $b_2^*$ "picks up" the effect of the omitted variable

The following figure demonstrates this connection

## *Picking Up the Effect of an Omitted Variable*

$$E(b_2^*) = \beta_2 + \beta_3 b_{23} \neq \beta_2$$



These results have important practical applications:

- If an estimated LS model has parameter estimates with unexpected signs or unrealistic magnitudes, a possible cause is the omission of an important variable

- Standard statistical tests are not appropriate since these tests require estimators that are unbiased

While omitting a variable from the regression usually biases the least squares estimator, the bias will be zero if

$$[\beta_3 = 0,\ \ b_{23} = 0, \text{ or both}]$$

Only under these conditions will the least squares estimator in the mis-specified model be unbiased

Finally, the bias caused by an omitted variable generalizes to the case where two or more independent variables are omitted

In this more general case, the bias of an estimator is a linear combination of the population parameters of the omitted variables, with the weights being the slopes from auxiliary regressions of each of the omitted relevant variables on the included variable

For a model with three independent variables and two of the three omitted

$$E(b_2^*) = \beta_2 + \beta_3 b_{23} + \beta_4 b_{24} \neq \beta_2$$

and

$$\text{Bias} = E(b_2^*) - \beta_2 = \beta_3 b_{23} + \beta_4 b_{24}$$

where

$b_{23}$ is the slope estimator from an auxiliary regression of $x_{2,t}$ on $x_{3,t}$

$b_{24}$ is the slope estimator from an auxiliary regression of $x_{2,t}$ on $x_{4,t}$

The potentially dire consequences of omitting relevant variables may lead you to think that a good strategy is to include as many variables as possible in your model

However it may inflate the variances of your estimates because of the presence of irrelevant variables

Let's return to the Bay Area Rapid Food model to explore the consequences including an irrelevant variable

As before, assume the true model is

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$$

where $y_t$ is total chain revenue for week $t$, $x_{2,t}$ is average price of chain products in week $t$, and $x_{3,t}$ is advertising expenditures for week $t$

Now, suppose we decide to include in the model a fourth variable, $x_{4,t}$, representing the unemployment rate in the Bay Area

So, we estimate the following model

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4,t} + e_t$$

However, $x_{4,t}$ is an irrelevant variable and in reality $\beta_4 = 0$

- Including $x_{4,t}$ does not make the least squares estimators of the parameters for the other variables biased

- Including $x_{4,t}$ does mean the sampling variances of $b_1$, $b_2$ and $b_3$ generally will be greater than those obtained by estimating the correct model

The inflation of sampling variances follows from the Gauss-Markov Theorem that LS estimators for the correct model are minimum variance linear unbiased estimators

Note that the inflation of sampling variances will not occur if the unemployment rate (irrelevant variable) is uncorrelated with price and advertising (relevant variables)

## Summary of Key Results for Omitted and Irrelevant Variables

If a relevant variable is omitted from a model

$\Rightarrow$LS parameter estimators generally will be biased, but sampling variances of the estimators will be reduced

If an irrelevant variable is included in a model

$\Rightarrow$LS parameter estimators generally will be unbiased, but sampling variances of the estimators will be inflated

***Fundamental tradeoff between bias and efficiency***

## Monte Carlo Simulation Experiment on Specification Errors

It is difficult to gain an intuitive understanding of the impact of specification errors

We will use Monte Carlo simulations where we know the correct model to discern the bias-sampling variance impact of omitting a relevant variable or including an irrelevant variable

First, assume the following linear demand model generates samples of data,

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$$

where $y_t$ is quantity demanded in year $t$, $x_{2,t}$ is the price of the commodity in year $t$ and $x_{3,t}$ is consumer disposable income in year $t$

Next, let's assume the true population parameters are

$$\beta_1 = 15 \qquad \beta_2 = -1.6 \qquad \beta_3 = 0.7 \qquad \sigma^2 = 16$$

Finally, assume a sample size of 20 and values for $x_{2,t}$ and $x_{3,t}$ as shown in the following table

***Table 9.2*** Treatment Values Chosen for **p**, **y**, and **ps**

| p | y | ps |
|---|---|---|
| 5.40 | 32.45 | 5.85 |
| 5.18 | 34.29 | 5.23 |
| 5.18 | 34.29 | 5.23 |
| 5.01 | 29.61 | 5.51 |
| 5.55 | 31.45 | 6.10 |
| 4.86 | 29.98 | 5.67 |
| 5.45 | 32.04 | 5.09 |
| 5.15 | 32.91 | 5.90 |
| 5.63 | 37.36 | 6.02 |
| 5.53 | 35.94 | 5.22 |
| 5.75 | 30.93 | 6.56 |
| 6.47 | 33.56 | 6.91 |
| 6.20 | 35.87 | 6.17 |
| 6.13 | 36.69 | 6.35 |
| 5.67 | 36.74 | 6.42 |
| 6.43 | 41.28 | 6.41 |
| 6.65 | 40.27 | 6.49 |
| 7.19 | 42.78 | 6.46 |
| 6.49 | 35.81 | 7.58 |
| 7.42 | 43.22 | 7.58 |
| 7.20 | 42.26 | 7.18 |

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

The assumptions of the Monte Carlo experiment can be summarized as,

- $y_t = 15 - 1.6x_{2,t} + 0.7x_{3,t} + e_t$
- $e_t \sim N(0, 16)$
- Sample size of $T=20$
- Values shown in table for $x_{2,t}$ and $x_{3,t}$

Based on these assumptions, we know that

$$\text{var}(b_1) = 60.14$$
$$\text{var}(b_2) = 5.44$$
$$\text{var}(b_3) = 0.17$$

and

$$b_1 \sim N(15, 60.14)$$
$$b_2 \sim N(-1.6, 5.44)$$
$$b_3 \sim N(0.7, 0.17)$$

## *Simulation I: Correct Model*

Data Generation Model: $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$

Model Estimated: $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$

1. Compute the expected value for $y_t$ at each level of $x_{2,t}$ and $x_{3,t}$ using the "true" population line of $y_t = 15 - 1.6x_{2,t} + 0.7x_{3,t}$

2. Randomly draw 20 error term values from the distribution $e_t \sim N(0, 16)$

3. Add the 20 random error values to the expected value for $y_t$ computed in step (1), which will create 20 "sample" observations on $y_t$

4. Regress the 20 "sample" observations of $y_t$ on the 20 values for $x_{2,t}$ and $x_{3,t}$

5. Save the estimates for $b_1$, $b_2$, $b_3$, $\hat{\sigma}^2$, $\hat{\text{var}}(b_1)$, $\hat{\text{var}}(b_2)$ and $\hat{\text{var}}(b_3)$

6. Repeat steps (2) - (5) 1,000 times

## *Simulation II: Omitted Variable Model*

Data Generation Model: $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$

Model Estimated: $y_t = \beta_1 + \beta_2 x_{2,t} + e_t$

1. Compute the expected value for $y_t$ at each level of $x_{2,t}$ and $x_{3,t}$ using the "true" population line of $y_t = 15 - 1.6 x_{2,t} + 0.7 x_{3,t}$

2. Randomly draw 20 error term values from the distribution $e_t \sim N(0, 16)$

3. Add the 20 random error values to the expected value for $y_t$ computed in step (1), which will create 20 "sample" observations on $y_t$

4. Regress the 20 "sample" observations of $y_t$ on the 20 values for $x_{2,t}$ [KEY CHANGE]

5. Save the estimates for $b_1$, $b_2$, $\hat{\sigma}^2$, $\hat{\text{var}}(b_1)$ and $\hat{\text{var}}(b_2)$

6. Repeat steps (2) - (5) 1,000 times

## *Simulation III: Irrelevant Variable Model*

Data Generation Model: $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$

Model Estimated: $y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4,t} + e_t$

1. Compute the expected value for $y_t$ at each level of $x_{2,t}$ and $x_{3,t}$ using the "true" population line of $y_t = 15 - 1.6 x_{2,t} + 0.7 x_{3,t}$

2. Randomly draw 20 error term values from the distribution $e_t \sim N(0, 16)$

3. Add the 20 random error values to the expected value for $y_t$ computed in step (1), which will create 20 "sample" observations on $y_t$

4. Regress the 20 "sample" observations of $y_t$ on the 20 values for $x_{2,t}$, $x_{3,t}$ and $x_{4,t}$ **[KEY CHANGE]**

5. Save the estimates for $b_1$, $b_2$, $b_3$, $b_4$, $\hat{\sigma}^2$, $\hat{\text{var}}(b_1)$, $\hat{\text{var}}(b_2)$, $\hat{\text{var}}(b_3)$ and $\hat{\text{var}}(b_4)$

6. Repeat steps (2) - (5) 1,000 times

***Table 9.3*** Sampling Experiment Results for Correct, Under-, and Overspecified Models

| Model | Mean from 1000 Samples | | | | Bias from 1000 Samples | | | | Variance from 1000 Samples | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ |
| Correct $(x_1, x_2, x_3)$ | 15.11 | −1.66 | 0.70 | 0 | 0.11 | −0.06 | 0 | 0 | 55.42 | 5.22 | 0.16 | 0 |
| Underspecified $(x_1, x_2)$ | 19.88 | 1.77 | 0 | 0 | 4.88 | 3.37 | −0.70 | 0 | 48.02 | 1.33 | 0 | 0 |
| Overspecified $(x_1, x_2, x_3, x_4)$ | 14.92 | −1.75 | 0.71 | 0.08 | −0.08 | −0.15 | 0.01 | 0.08 | 85.90 | 13.00 | 0.21 | 5.68 |

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

## Specification Uncertainty: What's a Researcher to Do?

Let's begin by stating a "hard" truth:

> Selecting the appropriate set of independent variables and the correct functional form is a difficult problem in econometrics that does not have a satisfactory solution

As a result, a common strategy in the presence of specification uncertainty is to:

1. Try a large number of combinations of independent variables and functional forms

2. Then report the single specification with the "best" statistical results (e.g., correct signs, high $t$-stats, large $R^2$) as if it were the only one estimated

Formally, this is called a "specification search" and it has potentially dire statistical problems!

## *Statistical Consequences of Specification Searching*

The sampling properties of the specification search "estimator" are unknown

- LS estimates of sampling variability are biased downwards, as the final model does not take into account the specification search in the first step
- Confidence intervals will be too narrow, test statistics will have inflated *p*-values and goodness-of-fit will be overstated
- Parameter estimates may be biased due to the omission of theoretically-relevant variables
- Out-of-sample forecasts may be quite inaccurate

Studenmund (2001):

***In other words, if enough alternatives are tried, the chances of obtaining results desired by the researcher are increased tremendously, but the final result is essentially worthless. The researcher hasn't found any scientific evidence to support the original hypothesis; rather, prior expectations were imposed on the data in a way that is essentially misleading.***

Leamer (1983):

***The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum likelihood estimation, in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of computer output the one thorn of a model he likes best, the one he chooses to portray as a rose.***

Leamer and Leonard (1983) provide an especially damaging comment below:

***Empirical results reported in economic journals are selected from a large set of estimated models... Since the process is well known to professional readers, the reported results are widely regarded to overstate the precision of the estimates, and probably to distort them as well. As a consequence, statistical analyses are either greatly discounted or completely ignored.***

Extensive specification searching is often described by the derisive terms,

- "data-mining"
- "fishing expedition"
- "data grubbing"
- "number-crunching"

Coase:

***If you torture the data long enough, Nature will confess***

Leamer:

***There are two things you are better off not watching in the making: sausage and econometric estimates***

Econometrics profession has split into three camps on the issue of specification searching

## *Average Economic Regression (AER)*

Start with a simple model suggested by economic theory and assume it is correctly specified

- If problems are observed, e.g., wrong signs, insignificant *t*-stats or implausible magnitudes, then proceed to "test up" to a more general model
- Include additional independent variables and try different functional forms
- Emphasis on "traditional" criteria for modeling success: correct signs, high $R^2$, significant *t*-statistics and regression *F*-statistics

This procedure may be prone to omitted variable bias in the initial specification, but reduces sampling variance of the estimators by keeping the number of variables to a minimum in the initial specification

A significant problem with this approach is that only the "final model" typically is reported, with no reference to the process used to arrive at the specification

***Test, Test, Test (TTT)***

Basically, just the opposite of the AER approach

Start with the most general model possible, and then proceed to "test down" to a more specific model

- Use a battery of diagnostic tests to determine if a specification is "congruent" with the data
- Essentially, test which independent variables can be excluded

This procedure is not likely to be subject to omitted variable bias in the initial specification, but sampling variances may be substantially inflated by the inclusion of a number of irrelevant variables

- Resulting lack of precision may make it difficult to make reliable conclusions about key estimates

- Sometimes derisively labeled the “kitchen sink” approach to econometrics

The TTT approach does have the considerable merit of making the “specification search” transparent

## *Sensitivity Analysis*

Starting point is the observation that the other two approaches (AER and TTT) may be sensitive to the initial model and the order in which tests are conducted

Begin the analysis by identifying the “family” of competing specifications based on:

- Different economic theories
- Previous modeling work

Ask how sensitive key results are to the different plausible specifications

- “Sturdy” estimates are reasonably consistent across specifications

- "Fragile" estimates vary widely across specifications

The sensitivity analysis can be based on sophisticated Bayesian methods or more *ad hoc* procedures

If all reasonable specifications are examined, then researcher should have knowledge of the range of results for key estimates

Criticism of this approach centers around the weighting of the different models; different researchers are likely to have different weighting functions (may end up right where you started in terms of AER vs. TTT!!)

## *Personal Editorial*

My own econometric philosophy is best described as a mixture of the average economic regression approach and "low-tech" sensitivity analysis

- The logic of AER is appealing to me, especially the emphasis on simplicity
- But, I also want to know the sensitivity of key estimates to alternative specifications
- I am wary of key estimates that are "fragile"

  Leamer (1983):

  ***…an inference is not believable if it is fragile, if it can be reversed by minor changes in assumptions. As consumers of research, we correctly reserve judgment on an inference until it stands up to a study of fragility, usually by other researchers advocating opposite opinions.***

You should study the different approaches and be prepared to defend your choice of approach

## *Selection Criteria*

Criteria that may be useful in specifying an econometric model include:

1. Economic theory and logic

    - Foundation and starting point for identifying independent variables and functional form
    - Think carefully about the prediction of theory regarding marginal effects
    - Think carefully about logical bounds for variables (e.g. crop yields must be positive)

2. Compatibility with *a priori* expectations and previous studies

    - Researcher should always ask whether estimates are compatible with expectations based on theory and previous studies
    - Results should be compared to those in previous studies, and differences considered and explained

3. Simplicity

- Simple models often are less sensitive to changing assumptions and data than more complex models (property of "robustness")
- Simple models are easier to explain to decision-makers
- Simple models are easier to estimate, update and manipulate
- Principle of parsimony: all else equal, the simpler model is preferred

4. Use of *t*-tests and *F*-tests

- The relevance of a particular variable or set of variables may be suggested by *t*-tests of a zero null for individual parameters or an *F*-test of a joint zero null for a set of parameters
- Researchers traditionally make heavy use of such tests in model specification

- Important to remember that there are two possible reasons for a test outcome that does not reject a zero null hypothesis:

  a) The corresponding variables do not influence *y* and can be excluded from the model

  b) The corresponding variables are important ones for inclusion in the model, but the data are not sufficiently "informative" to prove that the variables are important (poor data problem)

- Because the "insignificance" of a parameter estimate can be caused by (a) or (b), you must be cautious about discarding variables based on *t*-tests or *F*-tests

- You could be excluding an irrelevant variable, but you also could induce omitted-variable bias in the remaining parameter estimates

## 5. Maximizing $R^2$

- Since $R^2$ tells us the proportion of variation in the dependent variable explained by the variation in the independent variables, another way of choosing between model specification is to select the model with the highest $R^2$

- This will give us the model with the highest explanatory power

- Three circumstances when comparison of $R^2$ across models is not appropriate:

  a) The models have different functional forms for the dependent variables

  b) The models have different numbers of independent variables

  c) One of the models does not have an intercept

## 6. Maximizing adjusted $R^2$

- Basic logic is similar to that of maximizing $R^2$
- $\bar{R}^2$ provides a measure of explanatory power that is adjusted for the number of independent variables in the model specifications
- Kennedy (p.91)

  ***It is worth reiterating that searching for a high $R^2$ or $\bar{R}^2$ runs the real danger of finding, through perseverance, an equation that fits the data well, but is incorrect because it captures the accidental features of the particular data set at hand (called "capitalizing on chance") rather than on the true underlying relationship***

7. Use of formal specification tests

- These tests are based on the observation that differences in explanatory power of models may be due to chance rather than true differences in fit

- A number of specification tests have been proposed (see pp. 78-79 of Kennedy for a categorization)

- The RESET test (Regression Specification Error Test), designed to detect omitted variables and incorrect functional form, is an example

  Suppose that we have specified and estimated the regression model

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$$

  Let the predicted values of the $y_t$ be

$$\hat{y}_t = b_1 + b_2 x_{2,t} + b_3 x_{3,t}$$

Consider the following two artificial models

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \gamma_1 \hat{y}_t^2 + e_t$$

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \gamma_1 \hat{y}_t^2 + \gamma_2 \hat{y}_t^3 + e_t$$

- A test for mis-specification in the first artificial model is a test of $H_0 : \gamma_1 = 0$ against the alternative $H_1 : \gamma_1 \neq 0$.
- A test for mis-specification in the second artificial model is a test of $H_0 : \gamma_1 = \gamma_2 = 0$ against $H_1 : \gamma_1 \neq 0,\ \gamma_2 \neq 0$ or both
- Rejection of $H_0$ implies the original model is inadequate and can be improved
- A failure to reject $H_0$ says the test has not been able to detect any mis-specification

Overall, the general philosophy of the test is: If we can significantly improve the model by artificially including powers of the predictions of the model, then the original model must have been inadequate

$\Rightarrow$ If we have omitted variables, some of their effect may be picked up by the terms $\hat{y}_t^2$ and $\hat{y}_t^3$

$\Rightarrow$ If we have the wrong functional form, the polynomial approximation that includes $\hat{y}_t^2$ and $\hat{y}_t^3$ may improve the fit of the model

- Another type of specification test is based on the Akaike Information Criterion (AIC) and the Schwartz Criterion (SC)

  Basic idea is to adjust the regression sum of squares (SSR) by a factor to create an index of the fit of the regression

$$AIC = \ln(SSR/T) + 2K/T$$

$$SC = \ln(SSR/T) + \ln(T)K/T$$

AIC or SC can be used to compare two or more alternative specifications

Objective in modeling is to minimize AIC or SC

Thus, the regression with the higher AIC or SC is the one with specification error

- Comparison of RESET, AIC and SC

The RESET test does not require a researcher to define an alternative specification, while AIC and SC do

Makes RESET easier to use, but potentially less informative than AIC or SC

$\Rightarrow$ RESET is most useful as a general test for the existence of specification error

$\Rightarrow$ AIC and SC are most useful for comparing two or more alternative models for specification errors

## Final Thoughts on Specification

The criterion described in the previous section can give some indication of when a model is mis-specified and whether one set of independent variables is preferable to another

However, when evaluating model results produced by any model selection methodology, a healthy dose of skepticism is required

All selection methodologies, formal or informal, involve two steps:

i) Screening competing models based on selection criteria

ii) Application of least squares to the final model selected

As we noted in the section on specification searching, the sampling properties of this two-step estimation process are unknown

- As a result, estimation results reported by the computer look better than they actually are

- The LS estimates of sampling variability for the final model do not take into account the specification search in the first step

Consequently, if data are subjected to a large battery of tests and criteria, there is a high probability of selecting a model that "looks good" but is not consistent with the true data generating process (Back to specification searching!!)

***Where does this leave the applied researcher?***

In the end there is no universally satisfactory solution to the problem of model specification

It simply may be too much to ask of the data (nature) to do all three of the following tasks:

1) Provide estimates of the unknown parameters

2) Help to determine the correct set of explanatory variables

3) Help to determine the correct functional form

***Statistical theory only assumes the data have to do the first task!***

Despite this seemingly intractable problem, models still have to be specified and estimated in practice

A good researcher will use a judicious combination of imagination, creativity, economic theory, expertise regarding the data and model selection criteria

You must always be prepared to defend the procedure used in a way that tells a logical and believable data analysis story

Johnston (1984) provides some wise advice regarding econometric research in practice. He suggests researchers should:

- Talk with experts in the area being modeled
- Become familiar with the relevant institutions
- Actually look at the data
- Avoid data mining
- Use economic theory
- Exploit the judgment of an experienced critic

Studenmund (2001) suggests the following guidelines for an "Ethical Econometrician:"

***We think there are two reasonable goals for econometricians when estimating models:***

1. ***Run as few different specifications as possible while still attempting to avoid the major econometric problems.***

2. ***Report honestly the number and type of different specifications estimated so that readers of the research can evaluate how much weight to give your results.***

***Therefore, the art of econometrics boils down to attempting to find the best possible equation in the fewest possible number of regression runs. Only careful thinking and reading before estimating the first regression can bring this about. An ethical econometrician is honest and complete in reporting the different specifications and/or data sets used.***